



# IPTC Mirror

## Advancing the new News Architecture

**Refining the News Architecture is proving a complex and time-consuming task. Presentation of the NAR model and specification at the Autumn Meeting raised a number of issues, which the development group have been dealing with, and a formal NAR Release Candidate should be available for consideration early in 2007.**

Outlining the status of News Architecture (NAR) development, Working Party Chair Laurent Le Meur (AFP) explained that efforts were being concentrated on the production of a NAR Release Candidate, with the draft model and specification and associated XML Schema. The NewsContent Working Groups and the NewsCodes Working Party would then be asked to carry out an evaluation.

The second Experimental Phase - EP#2 - was designed to investigate how well the NAR performed as a generic model for the exchange of newsworthy information. This phase ran until the end of August and while the results were generally satisfactory a number of improvements to the specification were identified.

At the same time development work had been continuing and the resulting - revised and extended - version of the specification was made available to members prior to the Autumn Meeting.

### Development aims

Goals behind NAR development are to:

- Simplify the processing of news

objects.

- Manage news events, sports results and other news-related information in the same way.
- Maintain compatibility with the current NewsML 1 Model.
- Make it compact.
- Make it storage friendly.
- Leverage of semantic capabilities.
- Use of current XML technologies.

### Tension

However, the wide scope of these aims creates tension in the development process.

On one side there is the need for simplicity, to give a high level of usability and interoperability along with easy compatibility with web applications.

On the other hand the approach has to satisfy the needs of high-profile news providers with precise control of content creation and usage, and a rigorous structure and associated processing.

### Profiles

The solution adopted is to have a standard with a "core" profile that is easy to learn and easy to use, and a "power" profile - a superset of the

## IPTC Meetings 2007

### Spring Meeting

Cairo, Egypt

**12 - 14 March 2007**

Semiramis Intercontinental Hotel

*Bookings for the Spring Meeting will open in early January 2007 and must be completed by 29 January.*

### Annual General Meeting

Tokyo, Japan

**28 - 31 May 2007**

### Autumn Meeting

Prague, Czech Republic

**15 - 17 October 2007**

## Summary

**IPTC Meetings Schedule 2007 - Page 1.**

### NewsML Architecture Working Party

NAR development aims and challenges - **Page 1**.  
Knowledge management - **Page 2**. Further modifications carried out to meet specific requirements - **Page 2**. Development timeline proposed for the NAR and the G2-standards family - **Page 3**.

### NewsCodes Working Party

SchemaLogic Taxonomy Management system to be used for NewsCodes development and maintenance - **Page 2**.

New IPTC Members - **Page 2**.

News Domain Research Projects at Carlos III University, Madrid - **Page 4**.

The IPTC Mirror has hyperlinks for **web addresses** and for **page references**.

core profile with added features and provision for provider defined extensions. To ensure compatibility processing systems have to be able to deal with the main content and ignore unrecognised provider-specific extensions.

### Knowledge handling

Laurent went on to explain that one of the main challenges being addressed by the NAR was the need to combine news and knowledge. The current reality is extensive use of keywords but in future the aim was to extract the "information DNA" from text - this would involve the identification of concepts and entities that form the subjects of news - such as people, organisations, themes, and events - along with establishing relationships between the concepts.

### ConceptItem

For this the NAR provides the ConceptItem as a model for knowledge (in the same way as a model for news is provided by the NewsItem). Concepts may be named entities

(such as people or places) or generic concepts.

The ConceptItem makes it possible to manage and exchange information about a concept, and each ConceptItem has a unique identifier applied by the information provider. Content of a ConceptItem is typically a concept definition which consists of free text description and properties.

Specific properties have been defined for some common concepts. For example the "person" concept has properties of born, died, gender, affiliation, occupation, skills, and contact information.

### Concept Identifier

Different ConceptItems from different providers may contain information about the same concept, with a Concept Identifier being used to identify the concept itself.

Concept Identifiers have to be unambiguous but are not unique as there may be more than one identifier for a specific concept. They normally take the form of a QCode (Qualified Code) - essentially a

code within a scheme.

An important feature is the ability to specify relationships between concepts with specific provision being made for the common "sameas" "broader", "narrower" and "related" relationships.

### KnowledgeItem

To provide further flexibility in the use of concepts a new Item - the KnowledgeItem - has been introduced and is broadly comparable to the NewsML 1 TopicSet.

A KnowledgeItem consists of a set of concept definitions grouped together in a consistent structure, and can be managed and exchanged as a whole. A typical use example would be in the management and exchange of a controlled vocabulary.

### Evaluation

Prior to the Autumn Meeting it was intended that the draft of the News Architecture that had been circulated for the Meeting - draft 9 - would be the basis for a Release Candidate Version. With this in mind the Working Party approved a motion asking the News Content Working Groups to evaluate the new draft and provide written feedback.

It was hoped that this could take place over the period November 2006 to January 2007.

However, discussions at the Autumn Meeting - both during the working sessions and less formally - along with further input from the Content Standards Working Groups - raised a number of significant points that had to be dealt with before testing could start.

### Evolution

Since the Autumn Meeting concentrated efforts to deal with the points raised have resulted in significant

## NewsCodes Taxonomy Management

Development and management of the IPTC NewsCodes will now be carried out using the SchemaLogic system which was demonstrated at the Autumn Meeting - see IPTC Mirror No 136 November 2006 - with a three-year agreement for use of the system having been signed in early December.

The system will provide IPTC with a central location for semantic data (hosted by SchemaLogic) that members can access to allow collaborative development and maintenance of the taxonomies for the NewsCodes.

### Easy access

Details of the agreement were announced jointly by IPTC and SchemaLogic ([www.schemalogic.com](http://www.schemalogic.com)).

In the announcement IPTC Managing Director Michael Steidl explained "Our members are spread throughout the world, so it was important to have a hosted solution that would make it easy for everyone to access and contribute to. SchemaLogic's technology will allow all of our members the opportunity to participate in the taxonomy creation and management process."

### Proving ground

Jeff Dirks, President and Chief Executive Officer of SchemaLogic said "Working with the IPTC will provide us a great opportunity to demonstrate the flexibility and breadth of our hosted service. This will also serve as a proving ground for delivering taxonomy management through a hosted service model, and we are excited to be working with such a reputable organization. IPTC's members impact the efforts of most of the major media organizations in the world, and they need a tool that will truly help them to collaborate effectively and manage a large and evolving taxonomy. We appreciate their trust in our technology and look forward to making this a very successful and rewarding project for both parties."

evolution of the Model. A revised Release Candidate was under final consideration by the NAR development group in early December, while the corresponding XML

Schema was being prepared by the consultants.

A package for the release Candidate will be released for public consideration in early January 2007.

In addition a timeline for final development, approval and release of the NAR and of the first G2-Standards has been proposed, as shown below.

## NAR and G-2 Standards approval Timeline

### 1 January 2007

**NAR:** Release Candidate package is made available to the general public A final review phase is started.

### 30 January 2007

**NewsML-G2:** Release Candidate for the structure specification (only) made available to the general public.

### 21 February 2007

**NAR:** Release Candidate final comments due. This is the very final date to propose substantial changes to NAR v1.0. Any requests received after this date will be deferred to v1.1.

### 2 March 2007

**NewsML-G2:** Comments on the Release Candidate structure due.

### 14 March 2007

#### Spring Meeting - Cairo

**NAR:** Report from the NAR-dev group on proposed final changes to the NAR, final discussion on details only by the NAR Working Party. Only minor changes will be implemented in NAR v1.0, substantial changes will be deferred to v1.1 in 2008.

**NewsML-G2:** Discussion on the NewsML-G2/structure specification.

### 2 April 2007

**NAR:** NAR-dev group releases the **FINAL** version of the NAR v1.0 specs which have to be considered as completely frozen from this day on. XML Schema implementation starts.

### 25 April 2007

**NAR:** 1.0 XML Schema implementation finished.

### 7 May 2007

**NAR:** Package of V1.0 specifications and documentation.

**NewsML-G2:** Package of specifications and documentation. Both released to the IPTC members for a vote at the AGM.

### 30 May 2007

#### Annual General Meeting - Tokyo

**NAR:** Vote on approval of NAR v1.0 (only a vote, no discussion of any specification issues).

**NewsML-G2:** Vote on approval of NewsML-G2 structure.

**G2-Standards:** discussions on drafts, work on the specifications continues.

### 12 June 2007

**G2-Standards:** release of a draft specification for public review.

### 13 July 2007

**G2-Standards:** End of the public review phase.

### 27 August 2007

**G2-Standards:** Release Candidate available for a final review.

### 8 September 2007

**G2 Standards:** Final comments on the Release Candidates due, review closed. Final specifications and the XML Schema implementation under development.

### 24 September 2007

**G2-Standards:** Packages of v1.0 specifications and documentation released to the IPTC members for a vote at the Autumn Meeting.

### 17 October 2007

#### Autumn Meeting - Prague

**G2-Standards:** Vote on approval of G2-Standards v1.0.

# News Domain Research Projects

The Universidad Carlos III de Madrid was founded in 1989, and has some 17000 students with an academic staff of around 1300. The Telematics research department offers degrees in both Telecommunication Engineering and Computer Science (at Bachelor and Master levels), along with postgraduate courses. Some of the degrees are taught in English and the department has an active research facility covering a wide range of topics.

During the Autumn 2006 Meeting (in Madrid) Professor Luis Sánchez Fernández, provided details of some specific projects related to the news domain.

## El Periotrónica

One of the first developments was El Periotrónica, a customised syndication system using news content from several Spanish electronic newspapers. Users select their preferences according to the content categories, keywords and date ranges.

## Infomedia

Infomedia was designed to provide technical solutions to various issues in the development of electronic newspapers:

*Customisation* was dealt with by a dynamic extension to El Periotrónica which learns user interests according to the news items selected. *Multiplatform support* is achieved by producing XHTML, XHTML Basic and PDF from NewsML and NITF documents, using templates and dynamically generated XSLT stylesheets. *Business Model* is designed to provide flexible and customisable pricing strategies.

## Infoflex

Based on the Semantic Web and Web Services, InfoFlex is an architecture for undertaking distributed queries and appropriate news item recovery (for display to the user) from several, distributed, content providers.



Professor Luis Sánchez Fernández

## News Intelligence

The NEWS project

([www.news-project.com](http://www.news-project.com))

was undertaken to develop News Intelligence Technology for the Semantic Web. It is a EU funded project and along with Carlos III University the partners are the Spanish News Agency EFE, the Italian News Agency ANSA, DFKI (The German Research Center for Artificial Intelligence) and Ontology Ltd (Israel).

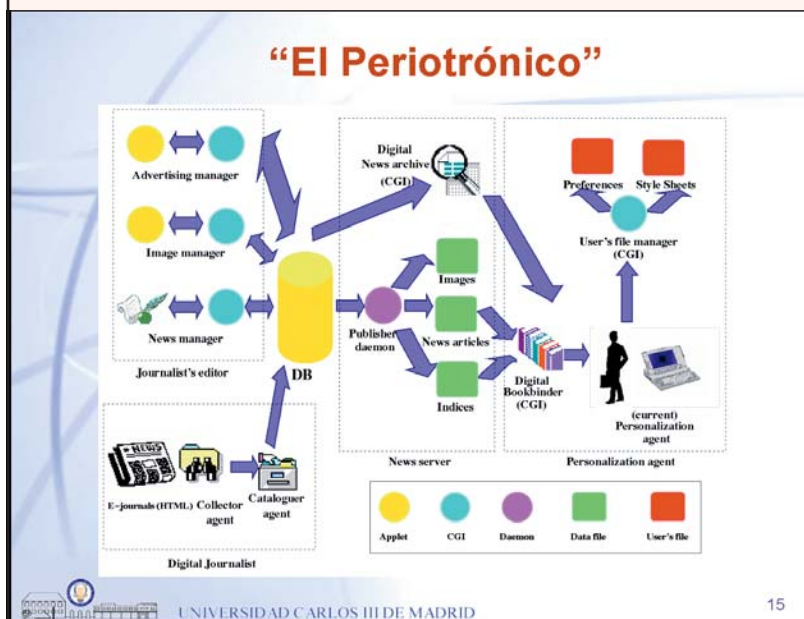
Components included in the system include: *RDF Schema Ontology* with a Management metadata module: *A categorisation module* (based on the IPTC Subject NewsCodes) and a content model based on the SUMO ontology. (SUMO - The Suggested Upper Merged Ontology - [www.ontologyportal.org](http://www.ontologyportal.org)) *Automated news item classification* using text analysis. Carried out in multiple languages it makes use of the IPTC Subject NewsCodes. *Automated content annotation* by entity extraction (such as person, organisation and location, though the coverage can be extended). This process uses a hybrid technology with morphological and

syntactical analysis and statistical analysis to give NewsCode categories and entity tagging.

*A heuristic and deductive database* allowing intelligent information retrieval. There is a relational database with a free text search engine and a reasoning element, and an entity identification system which maps entities to instances in the Ontology.

## Collaboration

In conclusion, Professor Fernández pointed out that the University have been working on the application of Web technologies to the news domain for about ten years. They find that the news domain is a good test area for their research activities and would be happy to undertake further collaboration with news agencies and the IPTC.



Published by the **International Press Telecommunications Council**  
 Royal Albert House, Sheet Street, Windsor, Berkshire SL4 1BE, England.  
 Managing Director: Michael W Steidl (mdirector@iptc.org). Editor: Hugh Johnstone (editor@iptc.org)  
 Tel: +44(0)1753 705051 Fax: +44(0)1753 831541 Web Site: [www.iptc.org](http://www.iptc.org)